Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 19 (2013) 155 – 162

Procedia Computer Science

The 4th International Conference on Ambient Systems, Networks and Technologies (ANT 2013)

Ad-centric Model Discovery for Predicting Ads' Click-through Rate

Zhe Gao[a], Qigang Gao[a]

[a]*Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada*

Abstract

Click here and insert your abstract text. Search engine advertising has become one of the most important revenue models of electronic commerce. It strongly affects the probability that users click on the ads at the side of the search results page if the system shows the right ones. To maximize the outcome of search engine revenue and improve users' perception on those ads, it is important to understand the factors which affect the click–through rate (CTR) on those ads. Tencent founded in 1998, is one of China's largest and most used Internet service portals. It provides a number of online services such as value-added Internet, mobile and telecom services and online advertising. As of September 30, 2011, Tencent had 711.7 million active Instant Messenger users. It forms the largest Internet Community in China. In this research, we use a very large dataset of Tencent click logs (soso.com) with millions records. First we describe how soso.com searching engine advertising works, our system architecture is designed with the click log dataset, and observations inside it aims at those ads with enough historical click logs. Then we show how to use ad-centric features to discover models that can find factors affecting CTR prediction performance. The proposed framework could help both optimizing the search engine system for soso.com and improving the ads designs for the advertisers.

Keywords: Sponsored Search, Large Dataset, System Design, CTR Prediction, User Behaviours, Term Factors, Regression Modeling;

1. Introduction

Sponsored research is one of the major revenue models of electronic commerce. Presenting ads directly on search results pages is a popular way of search engine advertising. In this research, we restrict to the pay-per-click (PPC) model which typically charges advertisers every time when their ads are clicked by users. To maximize revenue, PPC system must be able to predict users' browsing and click behaviour on ad results by similar features. For instance, if an ad was impressed 1000 times in the past and received 30

doi:10.1016/j.procs.2013.06.025

clicks, then the system could estimate its click-through rate (CTR) to be 0.03 and predict CTR of another ad having similar user or ads' features. Impression is how many times the ad results are returned and displayed to users when he or she had issued queries and CTR is only reasonably estimated based on a high impression of an ad.

In this research, we use log data from soso.com, a search engine service system of Tencent Company. Tencent, founded in 1998, is one of China's largest and most used Internet service portals. It provides some online services such as value-added Internet, mobile and telecom services and online advertising. As of September 30, 2011, Tencent had 711.7 million active Instant Messenger users and formed the largest Internet Community in China [1].

The dataset logged in the soso.com search engine gives massive click records for both new ads and those with many more impressions. Our goal is to build a model that could find general factors and predict the CTR of other ads using user related information and information about the ads themselves such as length of title, and other statistics information. It could help both optimizing the search engine's performance and improving ad design for the advertisers.

The research is based on Tencent's search engine (soso.com) online advertising system. A composite mark is calculated using both bids and quality of keywords. The score of keyword quality is affected by click history, ads and keywords, websites situation, relevance between ads and keywords and some other factors. And CTR is one of the most important terms to represent click history. The system will automatically make an anticipated ranking and the price is based on the counts of clicks by users. A bid for a keyword issued by a user is the highest price the user is willing to pay for a click on an ad produced by this keyword. This paper presents a framework how to effectively discover prediction models of CTR based on users' behaviour patterns from Tencent's very large click log dataset. The rest of the paper is organized as the following: related work is briefly surveyed in section 2. Section 3 introduces the system design and the implementation of the framework. Section 4 shows the results of experiments, and in section 5, we discuss the limitations and future work.

2. Related Work

Finding the most important features which affect users' click behaviour is crucial in predicting CTR. There has been much related research in the CTR prediction area. Richardson, Dominowska, and Ragno's [11] research focuses on using features of ads, terms, and advertisers to learn a logistic regression model in order to predict CTR for new ads that will improve the convergence and performance of an advertising system. In contrast, our system uses those ads having enough click log data. As shown in our experiments in Section 4, the CTR for total data is 3.488% while the CTR of the ads which have more than one impression is 5.429%. Thus, there should be differences in developing features and building models for both new ads and ads with historical click log data [11].

Kim, Qin, Liu, and Yu's [8] research aims at user clicks on ads from “the view of advertisers” [8]. They applied logistic function in a factor graph model to improve ad quality and advertising effectiveness, and then tried to find the factors that could explain user click behaviours. The relevance between query and ad is found as an important factor to explain user click behaviours. Besides, some words which are more attractive like “free” and “unlimited” causes more clicks. Although it is somewhat similar to our system, there are user-specified features to improve prediction performance. In Cheng and Cantu-Paz's [3] work, user-specific and demographic-based features were developed to reflect the click behaviour of individuals and groups with a maximum entropy model applied. However, it only works for users' information to improve the accuracy of CTR [3].

Dembczynski, Kotlowski, and Weiss's [5] work uses Decision Rules to predict ads' CTR. They defined ads with the same domain name as a target. The difference is that we could interpret the model to analyze and possibly improve it. It is not limited to the capability of CTR prediction, but more likely for making

more recommendations, hopefully improving the ads quality by those valuable interpretations [5]. In our system, interpretations are also applied to help advertisers to improve their ad design.

In data architecture, Linoff and Berry [9] mentioned the importance of abstraction level. The operational/Transaction data level is the base level which represents the base transactions of the systems [10]. Decision-support summary data is used for making decisions about the business by bringing new information integrated from raw source data. Metadata are used for business people to understand what database represents. For data miners, metadata provides valuable assistance in tracking down and understanding data [10]. "Business rules" is the top level of the abstraction representing the "relationships" in business [10].

Data flows through the data warehousing systems and will be finally transformed into the right information to end users. Linoff and Berry give the major components of data warehousing architecture [10]: Source systems, Extraction, transformation, and load (ETL), central repository, analytic sandbox, operational feedback systems, and end users. Analytic sandbox is the logic we used to support our research, and there are different types of ways to build analytic sandboxes [10]. As time goes by, although new technologies will merge and standard database will be enhanced by similar features, databases will not be affected by all technologies but only when optimizing SQL queries [10]. It is also more effective to use analytic functions directly in databases that could speed the algorithms. Analytic sandbox also supports statistical programming languages, while usually the processing works in databases [10]. What's more, MapReduce could be used to reduce the large quantity of data and Hadoop is a platform to support MapReduce process effectively, but it is not a requirement [10]. In our case, a simple random sampling is applied to reduce the data and statistics information is compared to confirm that the sampled data is good to use. In summary, we applied the ideas in our system building instead of using them directly.

3. Framework Design and Implementation

3.1. Model selection

Similar to Richardson et al's work [11]; we cast it as a regression problem as well, to predict CTR using related features. Related word usually uses logistic regression function:

$P = 1 / [1+e^{-(a_0+a_1X_1+a_2X_2+\ldots+a_nX_n)}]$.

In dataset, each record in the fact data represents an impression session instead of an impression. An impression session includes one or more than one impression with one or more than one click. So the logistic regression model does not fit our data. For the reason of effectively finding a global and interpretable pattern within a restricted time period, we adapt a multivariable linear regression model, and the target variable is:

$CTR = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n$.

To target on each ad instead of each impression session, it makes the process more flexible when building the model using data mining technique. By adding more and more variables, this equation could expand a best fit line equation. The independent variables are named 'X' in this equation and the value of each 'a' represents model coefficients (also called weight) to match those independent variables [9].

3.2. System Architecture

The system is mainly used for building a linear regression model to find factors affecting CTR prediction and predicting CTR. The aim is to design a flexible data processing system as shown in Figure 1. After importing click log data (around 10GB) into DBMS and creating indexes for queries, the total size exceeds the limitation of tablespace and the data processing on such a big table affects other DB users' experience. So, we decide to import them into workstations, a local desktop and a SSH connected

laptop as raw data storages. Moreover, queries are mostly executed in MySQL query browser and monitored in MySQL administrator, while some are executed in SPSS or Excel when stored in external files. Some data cleaning, reduction, and transformations are directly performed on the two workstations; for the purpose of flexibility, the cleaned, reduced and summarized data are exported into CSV files, analyzed in Weka, a data mining tool, and reimported into DB server again for later use. It is also convenience to use CSV file for analysis in SAS Enterprise Miner.

3.3. Click log data pre-processing

Each record of the raw dataset contains the following information: impression, click, query_id, keyword_id, user_id, ad_id, advertiser_id, title_id, description_id, display_url, depth, position. There are other 5 data files: User profile, Keyword, query, title, description. Each record of the last 4 files maps an id to a list of tokens. A token can basically be a word in a natural language.

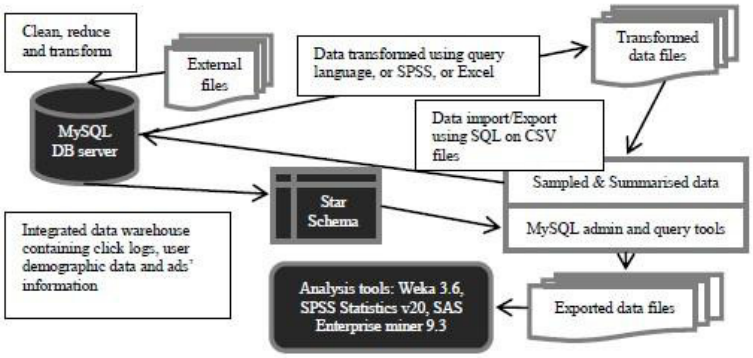

Fig. 1 The system architecture

Data summary, cleaning and reduction: The main dataset includes 14,847 advertisers with 641,707 ads. Each ad may have more than one keyword. There are totally 22,023,547users issued 24,122,076 queries with 235,582,879 ads impressions and 8,217,633 clicks. According to the objective to build an ad-centric model, we treat each ad as our target and estimate CTR for each of them. As Richardson states, for ads that have been impressed many times, the estimate is simply the binominal maximum likelihood estimation (MLE) using total # of clicks for each ad divided by total number of impressions [11]. Our experiments also take the average CTR distributions of each impression count# to show a high variance of CTR estimates which have high impression count. Thus, finding a relative true CTR is important [11].

Table 1 Ad-centric data summary

	statistic	Total dataset	Removed impression outliers	Removed CTR outliers	Random sampling 10%
Impressions for each ad	Minimum	1	304	304	304
	Maximum	1350400	1350400	1350400	661257
	Mean	233.189	2530.681	2500.515	2585.425
	StdDev	4096.575	14149.282	13929.433	13581.566
Click-through rate (CTR) estimate for each ad	Minimum	0	0	0	0
	Maximum	20.25	3.131	0.166	0.166
	Mean	0.053	0.047	0.039	0.039
	StdDev	0.153	0.065	0.031	0.031
Ads count		641707	52283	50618	5033

In [2], it was also mentioned, as an example, “an ad with a true CTR of 5% must be shown 1000 times before we are even 85% confident that our estimate is within 1% of the true CTR”. So, a moderate number

of impressions of each ad are necessary. Table 1 is a summary of impressions, clicks and CTR of the whole population and sampled data. The last row 'Ads count' shows a total number of ads for each dataset sample. Each column shows different statistic information for each sample. In column 'Removed impression outliers', we removed outliers with lower impression count# based on interquartile ranges and remove extreme CTR values based on the same algorithm. Thus, the data including the 50,618 ads information shown in the 'Removed CTR outliers' column is our experiment data. wE take a random sampling to select 10% from 50,618 ads to learn user click behaviours, as the impressions and CTR statistic information of the whole ads and sampled ads are similar.

3.4. Feature design (Data Transformation, Data Integration) and Cube design

According to the objective of building an ad-centric model, we are going to find advanced dimensions using summarized data.

Feature analysis: Features are designed based on the ad-centric model focusing on 4 groups: user, position, advertiser, and query-relevant. A simple demographic-CTR distribution is shown in Figure 2. Some specific groups of users has higher CTR in dark areas such as male under 18 and people older than 40. Therefore, it might be affected by having user demographic features. Position might also be an important factor for modeling because of the significant differences of CTR among different positions. CTR of ads at position 1 is much higher than other 2 in our experiments. Advertisers might be an important factor as well because of the brands or awareness especially when impression increases more and more. Besides the appearances of those ad lists, users might click the ads when they recognized that they were noticed of the advertisers sometimes.

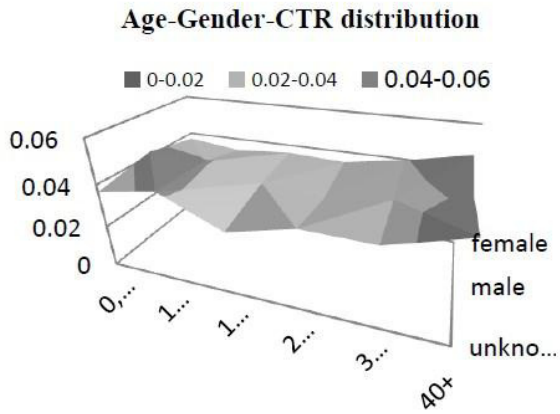

Fig. 2 Age-gender CTR distribution

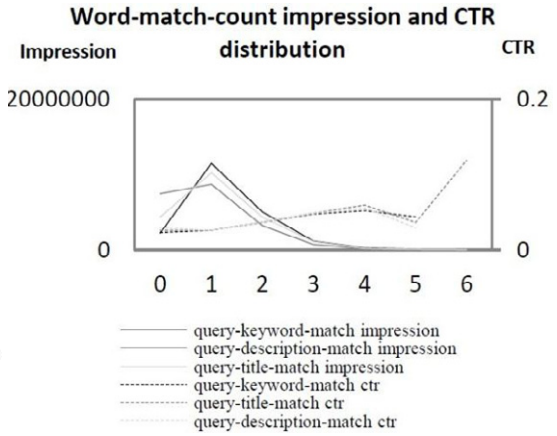

Fig. 3 Word match count impression and CTR distribution

In other words, the popularities of the ads might be reflected in the total count of ads issued by the same advertiser, and users may click more ads issued by the same advertiser. The matches between queries and ads' keywords, titles and descriptions are the most important part. It definitely affects users' decision. Moreover, Kim et al [8] mentioned some important words that could attract users more likely than other factors such as 'free' or 'unlimited'; users might have high possibility to click those ads. However, in the dataset, we have token_id as the identification instead of the word itself; we only consider the CTRs of those token_ids instead of the words. Figure 3 above is the CTR and impression distribution with how many words matched between queries and ads' title, description and keyword. It has an obvious increasing of CTR when words matched (X-axis) more. The tiny decreasing at word match count 4, 5 and much increasing at 6 occurs because of the instability of lower impressions.

Cube and feature design: As mentioned in section 3.4, dimensions are built based on the feature sets. A *star schema* is designed to be used for building data model. It contains 1 fact table and multiple dimension tables with calculated feature data. We also created advanced features as input variables of our regression model. Below are the lists of advanced features:

1. User-based features
- Ratio of impressions or clicks of different type of users in total clicks
- Most frequent age range and Standard deviation of users age most impressed

2. Position-based features
- Ratio of different positions' impressions/clicks in total
- Which position having most impressions/clicks

3. Advertiser-based features
- Distinct count ads/users/ users-clicked under each advertiser and their CTRs
- The ratio/count/ standard deviation of maximum number of different ads under the same advertiser was impressed/clicked to each user in count of all numbers of different ads

4. Query-ad relevant features
- Minimum/Maximum/Average/Mode query/title/description words/matched-words count for each advertisement
- The ratio of tokens which have higher CTR than the a's in title's/description's token count

4. Experiments and Evaluation

Table 2 Model R-square comparing

Models	R-square	Adj R-square
Model 1 including all features	0.8223	0.8173
Model 2 including advertiser-based features only	0.6985	0.6964
Model 3 including query-ad relevant features only	0.7511	0.7493

In regression tests, the data were divided into 60% for training and 40% for validation. We inferred the CTR for each ad using all features and trained other two models using advertiser-based features and ad-query relevant features separately. The P-values of the three models are all lower than 0.01%. It indicates that each of the equation is as a whole significant. However each feature set might affect CTR prediction, model 1 with all features applied is more stable as shown in R-squares tests in Table 2.

Table 3 model 1 output sample

Variable	Coefficient	T-score	P-value	Group
avg_desc_MIV_count_ratio	-0.0243	-18.3487	3.42E-71	query-ad relevant
user_countmaxadcount_click_ratio	-0.0352	-17.2599	1.38E-63	advertiser
advtr_ctr	0.2050	16.8790	5.13E-16	advertiser
user_countmaxadcount_imp_ratio	0.0183	9.2430	4.54E-20	advertiser
avg_title_MIV_count_ratio	-0.0102	-7.4826	9.64E-14	query-ad relevant
max_title_words_count	0.0009	-5.4372	5.87E-08	query-ad relevant
mode_query_words_count	-0.0010	-4.1206	3.89E-05	query_ad relevant
age std dev	-0.0135	-3.9793	0.0001	user
dist_click_user_ratio	-0.0054	-3.7257	0.0002	user
advtr_user_count	-3.23E-09	-3.4921	0.0005	advertiser
max_qd_click_word_mach_ratio	0.1768	3.2991	0.0010	query-ad relevant
max_qd_click_word_mach_count	-0.0283	-3.1620	0.0016	query-ad relevant
min_qd_click_word_mach_ratio	0.0050	3.0680	0.0022	query-ad relevant
min_qd_click_word_mach_count	-0.0042	-2.7850	0.0054	query-ad relevant
advtr_clicked_user_count	3.01E-08	2.6590	0.0079	advertiser

In each model, we can see all features' scores. By P-values in Table 3, top 15 features listed are all statistically significant for the model 1; By T-score, position-based features are not those relatively

important in the model 1 which include all features. However, we can't say it is truly not important as we have proved that the maximum likelihood estimated CTR of ads on different positions is clearly different. Some user-based features appears that female has more possibility to click on ads and shows a significant feature of users from 30 to 40 years old; but their lower T-score is because advertiser and query-ad relevant features are also included in this model which are much more important.

All models showed a strong correlation between CTR and query-ad relevant feature sets. However, we cannot initiatively say these features have only positive or negative effect on CTR prediction. For example, the average number of words in title (avg_title_words_count) has relatively negative effect on CTR prediction while the maximum number of words in title (max_title_words_count) has relatively positive effect on CTR prediction, and same others. If those related features are grouped to analyze, max_title_words_count has positive effect on CTR prediction, and avg_title_words_count is an adjustment of the effect from max_title_words_count. In all variables, 'title' also has another negative-value feature min_title_words_count to adjust the positive effect. Besides, another interesting finding is 'avg_title_MIV_count_ratio' which represents how many tokens having CTR larger than the ad's CTR. It strongly and negatively affects CTR prediction. Those tokens might be 'the', 'a' instead of 'free' or 'unlimited' according to the study of Kim, Qin, Liu, and Yu.

All complex feature set meet the same situation that it is not intuitionally saying these features has directly positive or negative effect on CTR prediction. Grouping by dimensions could apply. Richardson et al [11] mentioned that the best practice is to include as many features as possible although finding those top highly related features and overlapped features are important [11]. And we have showed that the model including all features is the best-fit one. Search engine could use the equation predict CTR and advertisers could also rank those features to improve their ads' design. Summarily in this case, factors list below in Table 4 are most important to predict CTR:

Table 4 most important factors

Most important factors
Tokens CTR distribution – ads attractiveness
How many ads issued by the same advertiser viewed or clicked by user,
How many ads issued by each advertiser - advertiser awareness
Length of title and description – ads attractiveness
Words matching between queries and ads' title and description
User demographic feature

Figure 4 presents the fit between targets and predicted CTR in percentile terms in validation dataset. The target mean CTR and predicted mean CTR are quite close. So the model is significant and is a good fit with the data.

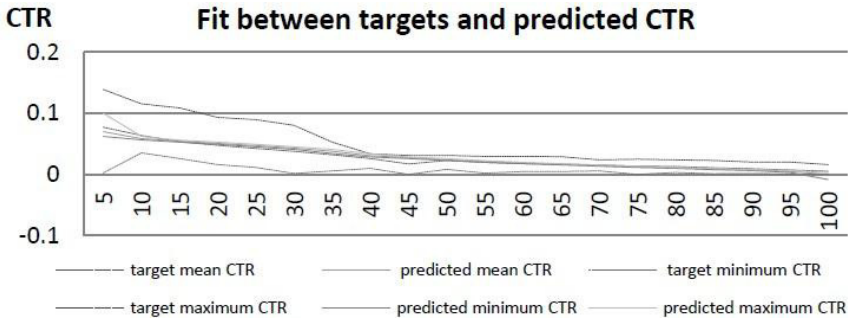

Fig. 4 The discovered fit between targets and predicted CTR

5. Conclusion

In this paper, we present a framework for model discovery to predict CTR and analyze factors using multivariable linear regression. Although there may be direct cause-and-effect relationship between the features and CTR prediction, some associations could still be interpreted among them. We can use grouping criteria to analyze the feature sets. It is a start of learning click behaviours in search engine advertising in a regression method, and we can extend researches into specific area in the future.

There are some limitations we met. First is the time and hardware performance limitations which cause us apply a random sampling method to effectively on querying on the dataset. This is also the reason why we use multivariable linear regression on an ad-centric model instead of logistic regression model, because there won't be a huge dataset with several records for each ad. Second is the dataset limitation. We were given a dataset without time series of each impression session. This causes us to focus on the appearances of the click actions on those ads. The dataset also lack of the words content but using the identifications instead. This causes the block to dive into the ads' content. For example, the attractiveness of ads requires the content to make analysis more accurate.

In the model result, we found that the features could predict results but some relationships among features and CTR are not understandable that we cannot have a deep interpretation. Especially when we look at the part of counting words matching. According to [14] research on user-click modeling, the task-centric click behaviour analysis could help to improve the accuracy of prediction. The session ends with satisfied search results on the premise of having time series of each impression and click [14]. For new ads, they don't have enough click behaviour, but could be clustered to predict CTR. What's more, the attractive analysis of ads is also interesting on the premise of having ads' text content, and the order of terms may also affect users' click behaviour. Deeply in advertisers' perspective, although CTR prediction is important, users may not buy anything even get into their commercial website. Advertisers could pay more attention on the link analysis between the clicks behaviour and buying behaviour.

6. References

[1] *About Tencent*. (2012). Retrieved from Tencent: http://www.tencent.com/zh-cn/at/abouttencent.shtml

[2] Burns, E. (2005, 09 22). *SEM Sees Optimization PPC*. Retrieved from Clickz: http://www.clickz.com/clickz/news/2138174/sem-sees-optimization-ppc

[3] Cheng, H., & Cantu-Paz, E. (2010). *Personalized CLick Prediction in Sponsored Search.* New York City.

[4] Chieh-Jen, W., & Hsin-Hsi, C. *Learning User Behaviors for Advertisements Click Prediction.*

[5] Dembczynski, K., Kotlowski, w., & Weiss, D. (2008). *Predicting Ads' Click-Through Rate with Decision Rules*. Beijing.

[6] Jiawei, H., & Micheline, K. (2006). Data Preprocessing. In M. K. Jiawei Han, *Data Mining: Concepts and Techniques* (pp. 47-104). San Francisco: Diane Cerra.

[7] Josh, A., Sandeep, P., & Torsten, S. (2009). *Modeling and Predicting User Behavior in Sponsored Search.* Paris.

[8] Kim, S., Qin, T., Liu, T.-Y., & Yu, H. (2011). *Advertiser-Centric Approach to Understand User Click Behavior in Sponsored Search.* Glasgow: CIKM'11.

[9] Linoff, G. S., & Berry, M. J. (2011). Data Mining Using Classic Statistical Techniques. In M. J. Gordon S. Linoff, *Data mining Techniques* (pp. 195-235). Indianapolis: Wiley Publishing Inc.

[10] Linoff, G. S., & Berry, M. J. (2011). Data Warehousing, OLAP, Analytic Sandboxes. In M. J. Gordon S. Linoff, *Data Mining Techniques* (pp. 613-654). Indianapolis: Wiley Publishing, Inc.

[11] Richardson, M., Dominowska, E., & Ragno, R. (2007). *Predicting Clicks: Estimating the Click-Through Rate for New Ads.* Alberta: IW3C2.

[12] Schwartz, B. (2010, 1 14). *Google AdWords Click Through Rates: 2% is Average But Double Digits is Great*. Retrieved from search engine round table : http://www.seroundtable.com/archives/021514.html

[13] *Tencent search advertising*. (2012). Retrieved from TQ365: http://soso.tq365.cn/

[14] Yuchen, Z., Weizhu, C., Dong, W., & Qiang, Y. (2011). *User-click Modeling for Understanding and predicting search-behavior.* San Diego.